International Journal of Research and Applications (July - Sep © 2014 Transactions) 1(3): 76-79
International Conference on Emerging Trends in Engineering & Technology (ICETET-2014)

ISSN (online): 2349-0020

ELECTRONICS ORIGINAL ARTICLE

Telugu Document Image Segmentation Methods

[1] Nagasudha D, [2] Madhaveelatha and [3] Y Pratap Reddy

[1] JNTU Manthani and [2 3] JNTU Hyderabad

ABSTRACT

The image segmentation is typically used to trace the object and boundaries such as line and curves in an image. The segmentation of the text reliability is necessary to perform the classification and Recognition. The main aim of segmentation is to partition the document image into various homogeneous regions such as text block, image block, line and word. Image segmentation is the front-stage processing of image compression. We hope that there are three advantages in image segmentation. The first is the speed. The second is good shape connectivity of its segmenting result. The third is good shape matching. Besides, we introduce many segmenting methods including iterative threshold technique, black pixels method and laplacian method for telugu document images.

Keywords: Edge Detection algorithm Preprocessing, Image acquisition. Image segmentation, histogram

I. INTRODUCTION

As the technology is enhancing in day to day life, there is huge amount of increase in the number of documents on the web. There is a need for an application that facilitates the user with an efficient retrieval of the information that is needed. Search engines are the key to finding specific information on the vast expanse of the World Wide Web. A search engine is a program that searches documents for specified keywords and returns a list of the documents where the keywords were found. The keywords to be searched for are given as query and the search engine gives the list of the documents having a match with the keywords in the query based on certain algorithms. The search engine also ranks the documents such that the more relevant documents are placed first in the results retrieved. For that segmentation of document images is to be done.

Modern technology has made it possible to produce, process, transmit and store digital images efficiently. Consequently the amount of visual information is increasing at an accelerating rate in many diverse application areas. The large amount of these image data are related to text. The information is stored in the form of digital versions and in document management system. Document image retrieval systems are utilized in many organizations which are using document image databases extensively. To fully exploit this different document image segmentation techniques are used.

When we refer to a paper document it is distinguished by the fact that it is on paper. However the notion of digital document is one which digital systems can understand and present to the user in an articulated manner. There are several types of documents which present information to a person that can be conceived and comprehended. Documents can be primarily divided into three different categories and they are Online: Documents that fall into the online paradigm consist of online handwriting that consist handwriting data captured by a digitizer that captures handwriting of a writer. These digitizers are specialized devices that capture a

writer' ink information his speed the pressure applied etc. which can be later used for further processing.

Offline: The least common denominator for handwriting is the paper and pen. Offline documents we do not access to information such the speed or pressure with which the writer must have written. Handwriting data in both online or offline could have cursive, discrete or mixed. Handwritten notes are examples of offline documents.

Printed: Printed documents contain textual information which is scanned copies from a book. The textual content in books is in the printed form following a specific font style font analysis. Such processing includes thresholding to reduce a gray scale or color image to binary image, reduction of noise to reduce extraneous data and thinning and region detection to enable easier subsequent detection of pertinent features and objects of interest, size and maintaining a standard uniformity across all pages. These are typically referred to as document images and could also contain images or pictures in addition to textual content. OCRs are used to extract the textual content from these documents, Books, Journals, articles, news papers; magazines are some of the examples of a printed document. Some of the popular digitizers for both offline and printed documents include the digital cameras, hand held and flat bed scanners etc.

Segmentation is to subdivide an image into its component regions or objects. It should stop when the objects of interest in an application have been isolated. Segmentation algorithms generally are based on one of 2 basis properties of intensity values

Discontinuity: to partition an image based on sharp changes in intensity (such as edges)

Similarity: to partition an image into regions those are similar according to a set of predefined criteria.

More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image (edge detection). Each of the pixels in a region is similar with respect to some characteristic or computed property, such as color or intensity, or texture. Adjacent regions are significantly different with respect to the same characteristic(s). When applied to a stack of images, typical in medical imaging, the resulting contours after image segmentation can be used to create 3D reconstructions with the help of interpolation algorithms like marching cubes.

The first segmentation method is Mean Gray Level., Edge Pixels, Iterative Selection and Percentage of Black Pixels.

Mean Gray Level: Mean Gray Level Algorithm is simply applied by summing up all the pixel values in the image and then taking the mean of it to obtain the threshold.

Percentage of Black Pixels: Assuming that percentage of black pixels is a constant for some types of images, lower pixel values up to the number of assumed pixels are segmented as background or black.

Edge pixels: Laplacian is calculated for each pixel and then histogram of pixels with large laplacians is created. Using this new histogram a threshold can be detected using any of the previous methods.

Iterative thresholding: In this method a threshold is iteratively calculated and refined by consecutive passes through the image.

The simplest method of image segmentation is called the thresholding method. This method is based on a clip-level (or a threshold value) to turn a gray-scale image into a binary image. There is also balanced histogram thresholding.

The key of this method is to select the threshold value (or values when multiple-levels are selected). Several popular methods are used in industry including the maximum entropy method, Otsu's method maximum variance), and k-means clustering.

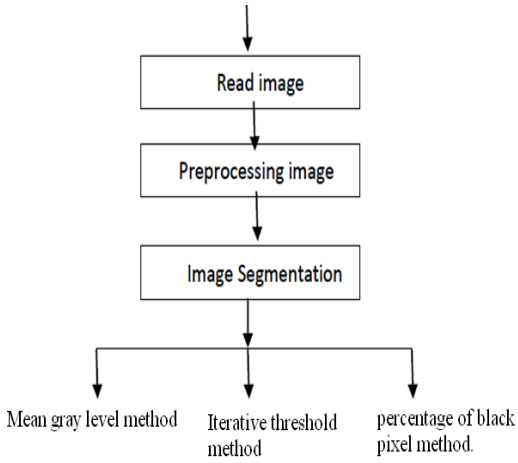
Histogram -based methods are very efficient when compared to other image segmentation methods

because they typically require only one pass through the pixels. In this technique, a histogram is computed from all of the pixels in the image, and the peaks and valleys in the histogram are used to locate the clusters in the image. Color or intensity can be used as the measure.

A refinement of this technique is to recursively apply the histogram-seeking method to clusters in the image in order to divide them into smaller clusters. This is repeated with smaller and smaller clusters until no more clusters are formed. One disadvantage of the histogram-seeking method is that it may be difficult to identify significant peaks and valleys in the image.

Histogram-based approaches can also be quickly adapted to occur over multiple frames, while maintaining their single pass efficiency. The histogram can be done in multiple fashions when multiple frames are considered. The same approach that is taken with one frame can be applied to multiple, and after the results are merged, peaks and valleys that were previously difficult to identify are more likely to be distinguishable. The histogram can also be applied on a per pixel basis where the information result is used to determine the most frequent color for the pixel location. This approach segments based on active objects and a static environment, resulting in a different type of segmentation useful in video tracking.

Edge detection is a well-developed field on its own within image processing. Region boundaries and edges are closely related, since there is often a sharp adjustment in intensity at the region boundaries. Edge detection techniques have therefore been used as the base of another segmentation technique. The edges identified by edge detection are often disconnected. To segment an object from an image however, one needs closed region boundaries. The desired edges are the boundaries between such objects or spatial axons. Segmentation methods can also be applied to edges obtained from edge detectors. Region growing methods mainly rely on the assumption that the neighbouring pixels within one region have similar values. The common procedure is to compare one pixel with its neighbours. If a similarity criterion is satisfied, the pixel can be set belong to the cluster as one or more of its neighbours. The selection of the similarity criterion is significant and the results are influenced by noise in all instances. We can conquer the noise problem easily by using some mask to filter the holes or outlier. Therefore, the problem of noise actually does not exist. In conclusion, it is obvious that the most serious problem of region growing is the power and time consuming

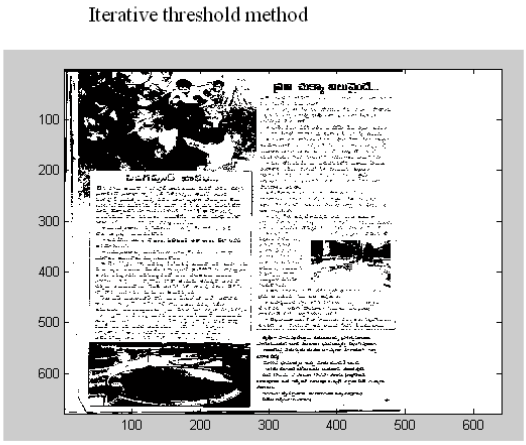


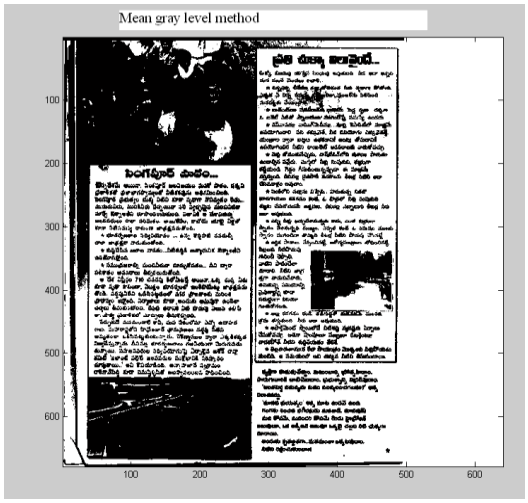
Algorithm: Iterative thresholding

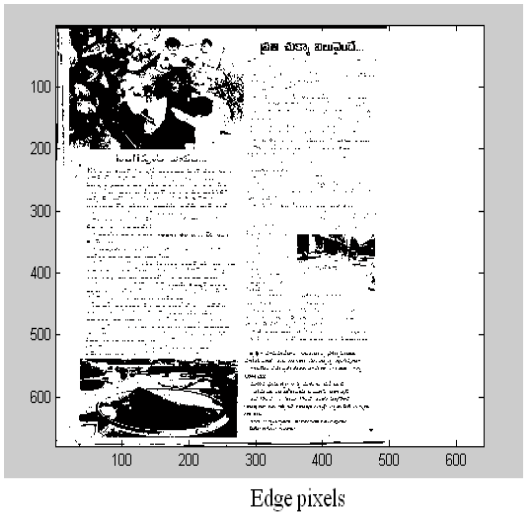
1. Assume that the four corner points of the image are background pixels (part of segment 0), and set $\mu 0$ to the average grey value of these four pixels. Assume all of the other pixels are object pixels, and set $\mu 1$ to their average grey value.
2. Set the threshold t to $t = 1$
2 ($\mu 0 + \mu 1$), and segment the image.

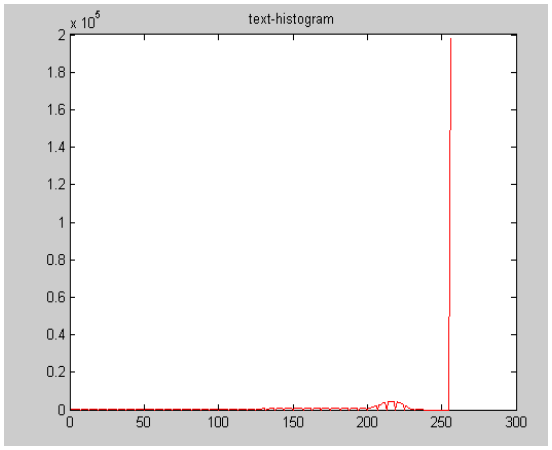
3. Recomputed $\mu 0$ and $\mu 1$, the mean of the original grey values of the two segments.

4. Go to step 2 and iterate until the threshold value no longer changes (or no longer changes significantly).









Conclusions

In this paper, telugu document image are segmented by using iterative threshold method, Mean gray level, detection method, laplacian method and percentage of black pixels are proposed.

References:

[1] K.H. Liang and J.J.W Mao, "Image Thresholding by Minimizing the Measures of Fuzziness", Pattern Recognition, Vol.28, No.1, PP.41-51, 1995.

[2] F. Samopa, A. Asano.,"Hybrid Image Thresholding Method using Edge Detection", IJCSNS International Journal of Computer Science and Network Security, Vol.9 No.4, PP.292-299, April 2009.

[3] Vijay kumar and Pankaj K. Sengar, "Segmentation of Printed Text in Devanagari Script and Gurmukhi Script", International Journal of Computer Applications,Vol 3, No. 8, June 2010,pp. 24-29

[4]N. Dhamayanthi, and P. Thangavel," Handwritten Tamil character recognition using neural network", Proceeding of Tamil Internet 2000, Singapore, July 22-24, 2000, pp. 171-176.

[6]Laurence Likforman-Sulem, et. al," Text Line Segmentation of Historical Documents: a survey", Submitted to Special Issue on Analysis of Historical Document, International Journal on Document Analysis and Recognition, Springer, 2006.